Coarse-to-Fine Classification and Scene Analysis

Donald Geman
Department of Applied Mathematics and Statistics
Center of Imaging Science, Whitaker Institute
Johns Hopkins University

Joint Work with **Sachin Gangaputra** and **Gilles Blanchard**

Outline

- Semantic Scene Interpretation
- A Statistical Framework for CTF Classification
  - Part I: Exploring a Hierarchy: “20Q Theory”
  - Part II: Constructing a Hierarchy
  - Part III: Assigning Likelihoods: The “Trace Model”

Semantic Scene Interpretation

- *Understanding how brains interpret sensory data, or computers might, is a major challenge*.
- Assume:
  - One grey-level image I. (Although cues from *color, motion or depth* are likely crucial to recognition.)
  - There is objective reality Y(I), at least at the level of key words.

Confounding Factors

- Local *(but not global)* ambiguity
- Arbitrary views and lighting
- Dominating clutter
- Infinite-dimensional classification

and ...

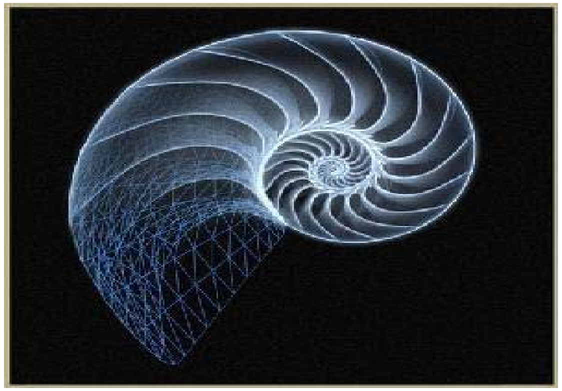
Three Dilemmas

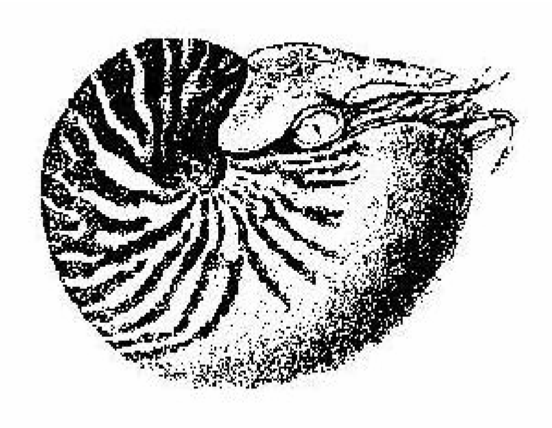
- Small Samples
- Bias vs. Variance
- Invariance vs. Selectivity

Detecting Boats

Where Are the Faces? Whose?

Within Class Variability





How Many Samples are Necessary?

Recognizing Context

Dreaming

A description machine

$$f : \mathbf{I} \rightarrow \mathbf{Y}$$

from an image $I \in \mathbf{I}$ to a description $Y \in \mathbf{Y}$ of the underlying scene.

Better Yet: A sequence of increasingly fine interpretations $Y = (Y_1, Y_2, ...)$, perhaps "nested."

Organizing Principles

- *Discrimination*: Proceed (almost) directly from data I to decision boundaries.
- *Data Generation*: Construct a joint statistical model for (features of) images I and interpretations Y.
- *Efficiency*: Exploit shared components among objects and interpretations to search for many things at once.

Efficiency-Driven Perception

- *Efficient representation, discrimination and computation all result from exploiting common "parts" and sub-interpretations.*

- Examples:
  - Compositional vision: A "theory of reusable parts"
  - Hierarchies of image patches or fragments
  - Coarse-to-fine classification

Outline

- Semantic Scene Interpretation
- A Statistical Framework for CTF Classification
  - Part I: Exploring a Hierarchy: “20Q Theory”
  - Part II: Constructing a Hierarchy
  - Part III: Assigning Likelihoods: The “Trace Model”

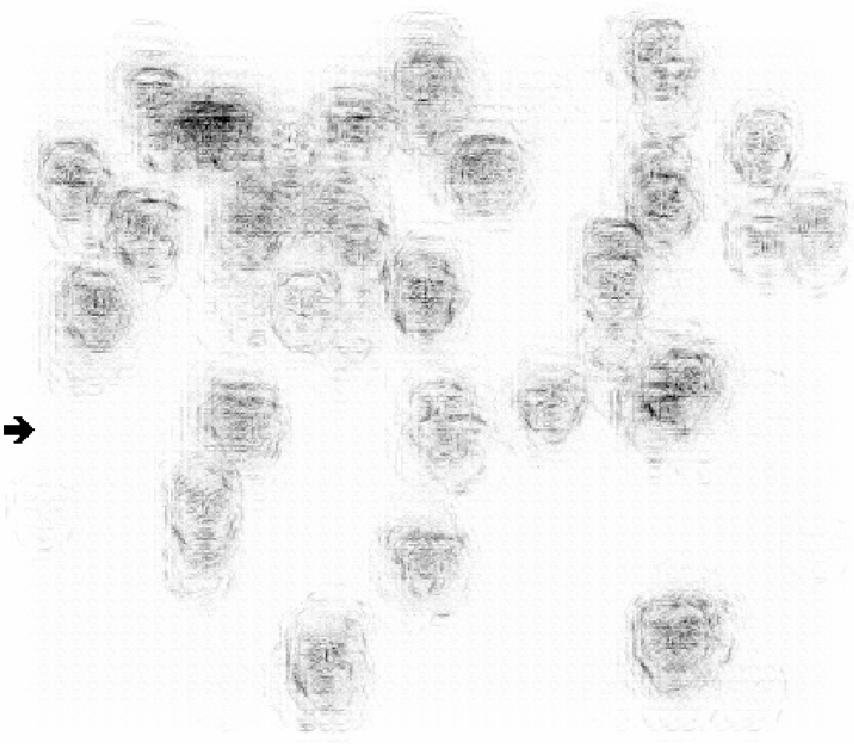
CTF Classification

Coarse-to-fine modeling of *both* the interpretations *and* the computational process:

- Unites representation and processing.
- Concentrates processing on ambiguous areas.
- Evidence that coarse information is conveyed earlier than fine information in neural responses to visual stimuli.

Density of Work

Original image



Spatial concentration of processing

Statistical Framework

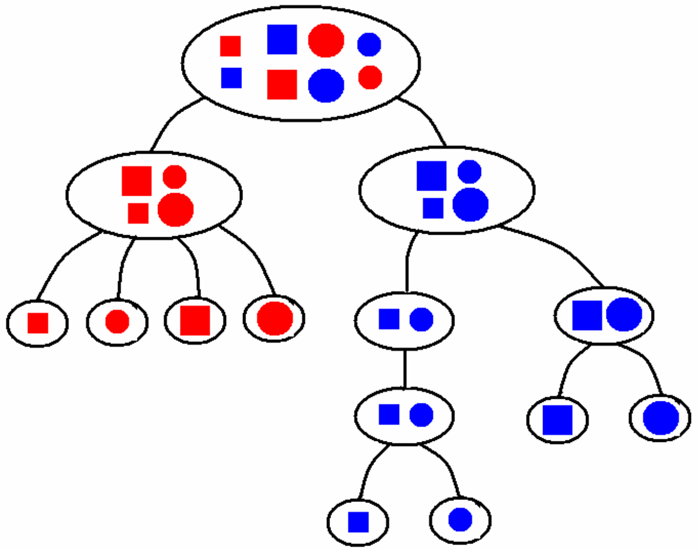
- There are natural groupings $A \subset \boldsymbol{Y}$ corresponding to "attributes"

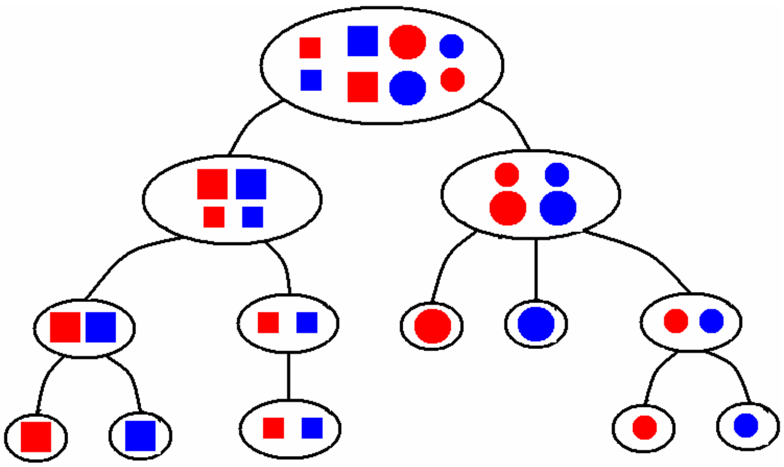
- In fact, there are natural nested partitions or hierarchies of attributes

$$H_{attr} = \{ A_{\xi},\ \xi \in T \}$$

where T is a tree graph.

Example: Attribute Hierarchies

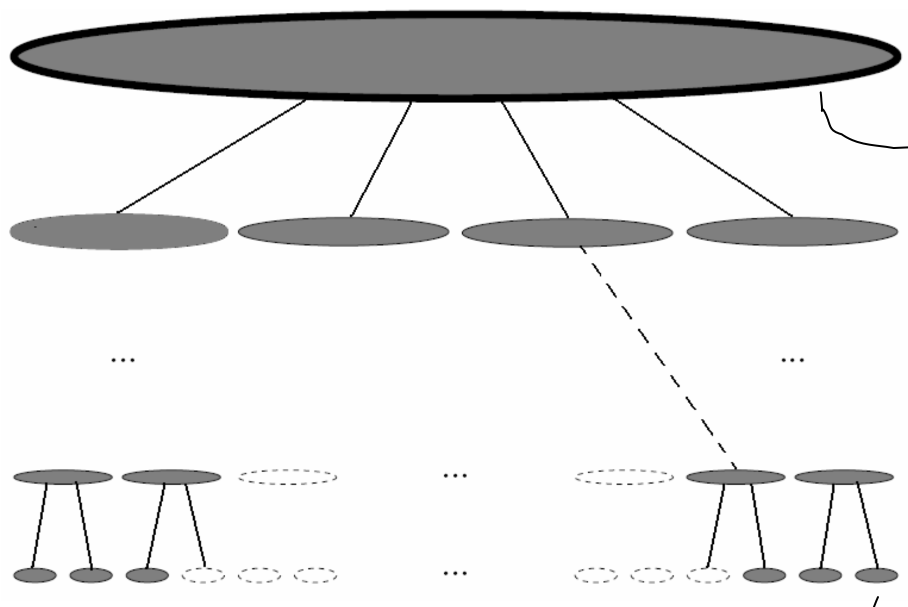




Example: Face Detection

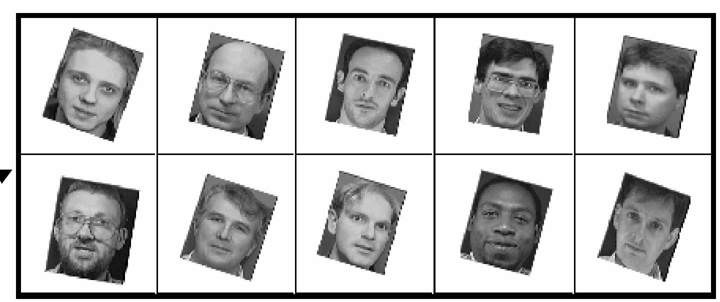
- I = subimage W (64x64 region)
- $\boldsymbol{Y} = \{(z,\sigma,\phi): z \in 8x8,\ 8 \le \sigma \le 15,\ -20^0 \le \phi \le 20^0\}$
- H_{att}: Constructed by considering 4 possible partitions for each "pose cell" A:
  - Quaternary split in location
  - Binary split in scale or orientation
  - No split (cascade)

Example: Pose Space



$\{ (z,\sigma,\phi)$: $z \in 8x8$,
$8 \leq \sigma \leq 15$, $-20^0 \leq \phi \leq 20^0 \}$

$\{ (z,\sigma,\phi)$: $z \in 2x2$
$14 \leq \sigma \leq 15$, $10^0 \leq \phi \leq 20^0 \}$



Statistical Framework (cont)

- For each $\xi \in T$, consider a binary test $X_\xi = X_{A_\xi}$ dedicated to $H_0{:}Y \in A_\xi$ against $H_a{:}\ B_{alt(\xi)} \subset \{Y \notin A_\xi\}$

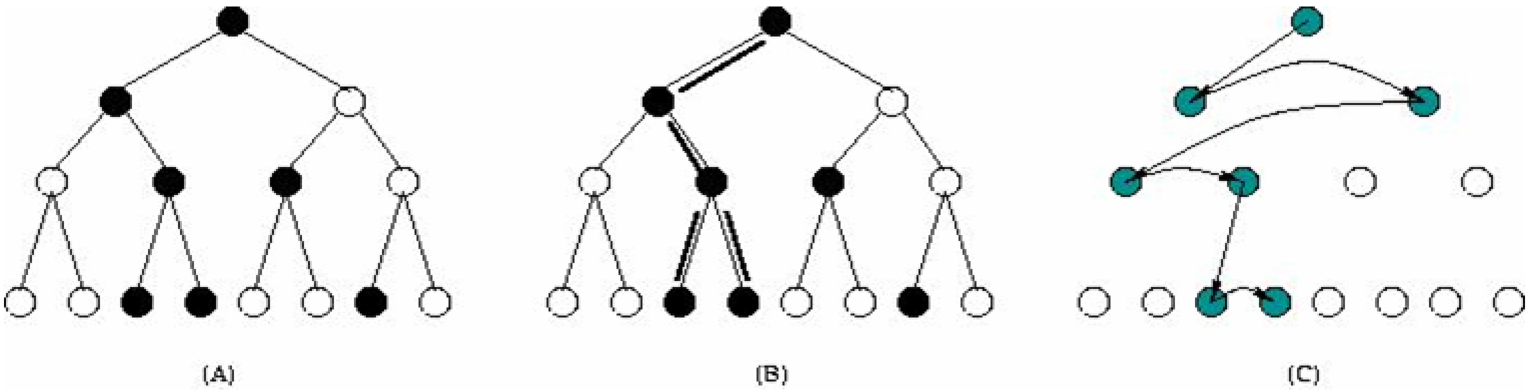
- Estimate Y by exploring $H_{test} = \{ X_\xi,\ \xi \in T \}$

  Constraint: Each X_ξ has a null false negative rate.

- Detections D: Explanations $y \in \mathbf{Y}$ not ruled out by any (performed) test:

$$D = \{y \in \mathbf{Y} :\ X_{A_\xi} = 1 \ \text{ for every } \xi \text{ such that } y \in A_\xi\}$$

Example



- A recursive partitioning of ***Y*** with four levels; there is a binary test for each of the 15 cells.
- (A): Positive tests are shown in black.
- (B): *D* is the union of leaves 3 and 4.
- (C): Tests performed under coarse-to-fine search.

Outline

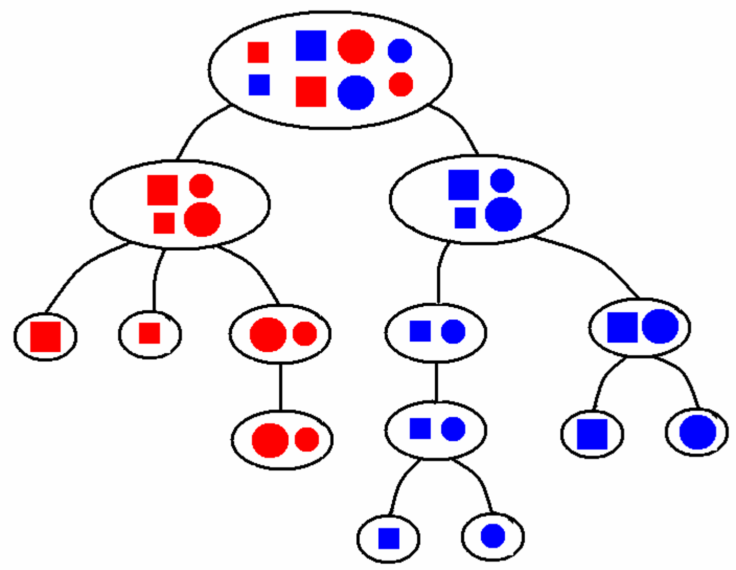
- Semantic Scene Interpretation
- A Statistical Framework for CTF Classification
  - Part I: Exploring a Hierarchy: "20Q Theory"
  - Part II: Constructing a Hierarchy
  - Part III: Assigning Likelihoods: The "Trace Model"

Part I: A 20Q Theory

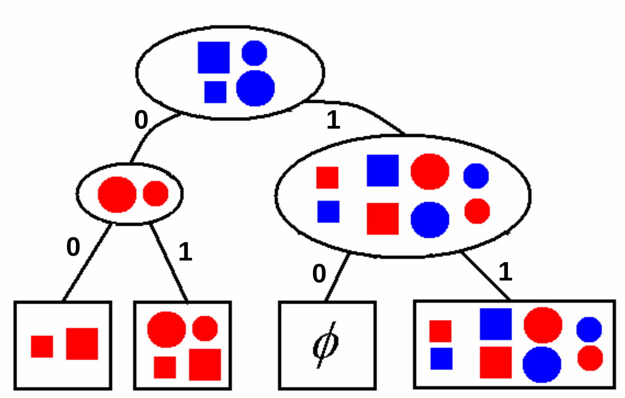
- Strategy: Adaptive (tree-structured) testing procedure:
  - $s \in S^0 \rightarrow X_{\xi(s)}$
  - $s \in \partial S \rightarrow \hat{Y}(s)$, the surviving explanations after testing.

- Cost: $c(X_\xi)$
- Power: $\beta(X_\xi) = P(X_\xi = 0 | B_{alt(A_\xi)})$

Representation vs. Processing



Representation tree



Decision tree representing a testing strategy

Computational Cost

- *Cost of Testing:* The sum of the costs before reaching a decision:

$$C_{test}(S) = \sum_{s \in \partial S} I_{H_s} \sum_{r \downarrow s} c(X_{\xi(r)})$$

$$E[C_{test}(S)] = \sum_{s \in S^0} c(X_{\xi(s)}) P(H_s) = \sum_{\xi \in T} c(X_\xi) q_\xi(S)$$

  where $q_\xi(S)$ is the probability of performing test X_ξ under the strategy S. H_s is the event node s is reached.

- *Total Computation:* $E[C_{test}(S) + c^* |\hat{Y}(S)|]$

Optimization

- *When are the strategies which minimize total computation CTF, meaning:*
  - $|A|\downarrow$ A monotonic decrease in scope.
  - $\beta \uparrow$ A monotonic increase in power

- *Two Fundamental Assumptions*:
  - Background domination: Take $P=P_0=P(.|Y=0)$ for measuring power and mean computation.
  - Conditional independence: The tests for distinct sets in H_{test} are independent under P_0

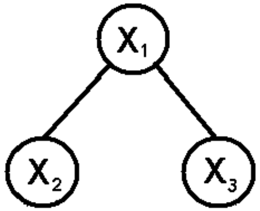
CTF Optimality Criterion

THEOREM: (G. Blanchard/DG) *CTF is optimal if*

$$\forall \xi \in T, \quad \frac{c(X_\xi)}{\beta(X_\xi)} \leq \sum_{\eta \in \mathcal{C}(\xi)} \frac{c(X_\eta)}{\beta(X_\eta)}$$

where $C(\xi)$ = direct children of ξ in T.

- *A numerical example:*



$c(X_1)=c(X_2)=c(X_3)$

$\beta(X_1) = 1/2, \ \beta(X_2)= \beta(X_3)=9/10$

Do X_1 first !

Outline

- Semantic Scene Interpretation
- A Statistical Framework for CTF Classification
  - Part I: Exploring a Hierarchy: “20Q Theory”
  - Part II: Constructing a Hierarchy
  - Part III: Assigning Likelihoods: The “Trace Model”

Part II: Hierarchy Design

- *Goal:* Construct the hierarchy and the tests simultaneously from training data
- Assume a universal learning algorithm

$$(A, \mathcal{L}) \rightarrow X_A$$

with $\alpha(X_A) = P(X_A = 0 | Y \in A) = 0$

- $\mathcal{L} = \mathcal{L}_+ \cup \mathcal{L}_-$ represents training examples
  - $\mathcal{L}_+ \sim \{ Y \in A \}$
  - $\mathcal{L}_- \sim B_{alt(A)} \subset \{ Y \notin A \}$

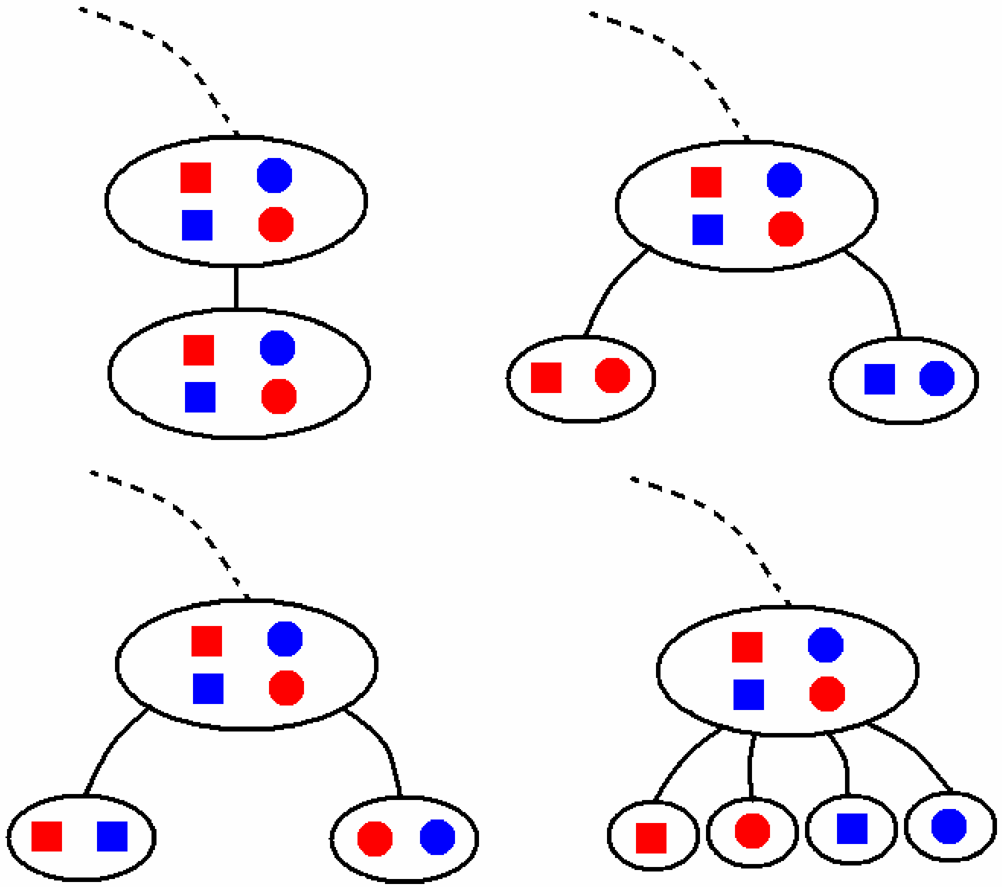
“Right” Alternative Hypothesis for CTF Search

- Alternate hypothesis at ξ: Conditional distribution of the data given $Y \notin A_\xi$ and the test X_ξ is performed. Due to CTF search X_ξ is performed $\Leftrightarrow$ all ancestor tests are performed and are positive:

$$B_{alt(\xi)} = \{Y \notin A_\xi\} \cap \{X_\eta = 1 \; \forall \; \eta \in \mathfrak{A}(\xi)\}$$

where $\mathfrak{A}(\xi)$ = ancestors of node ξ in T.

Which Decomposition?



Hierarchy Design (cont)

- Let $\Lambda(A)=\{A_1,A_2,\ldots,A_n\}$ *denote a* partition of A

- Combined test for $\Lambda(A)$:

$$X_\Lambda = \begin{cases} 1 \text{ if } X_{A_i} = 1 \text{ for some } i \\ 0 \text{ otherwise} \end{cases}$$

- *Cost* $c(X_\Lambda)=\sum_i c(X_{A_i})$
- *Power* $\beta(X_\Lambda) = P(X_\Lambda=0|B_{alt(A)})$

$$= P(X_{A_i}=0,\ i=1,\ldots,n \mid B_{alt(A)})$$

Hierarchy Design (cont)

- Given partitions $\Lambda_1, \Lambda_2, \ldots, \Lambda_k$ of A, choose:

$$i^* = \arg\min_{1 \le i \le k} \frac{c(X_{\Lambda_i})}{\beta(X_{\Lambda_i})}$$

- Now split A into $|\Lambda_{i*}|$ children and add these attributes to H_{attr} and the corresponding tests to H_{test}.

Special Case

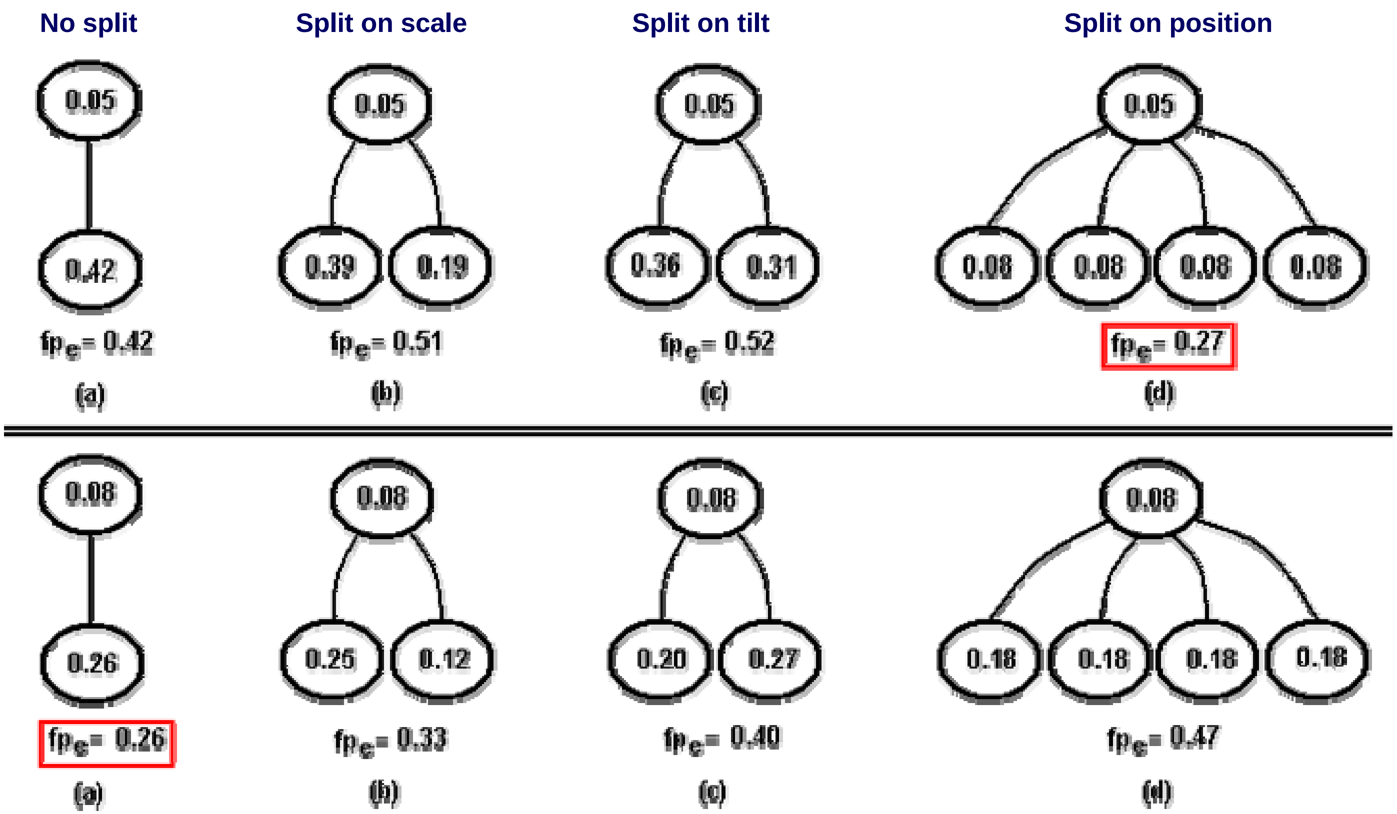
- Suppose $c(X_{\Lambda_i}) \equiv c$. For example,

$$c(A) \propto |A| \text{ for every } A \subset Y \text{ so that } c(X_{\Lambda_i}) \equiv |A|$$

- Then i^* is the partition which minimizes the false positive rate (per unit cost).

- *Recursive construction* of H_{attr}: Select the node with the highest false positive rate. Choose the split that minimizes the new (estimated) false positive rate.

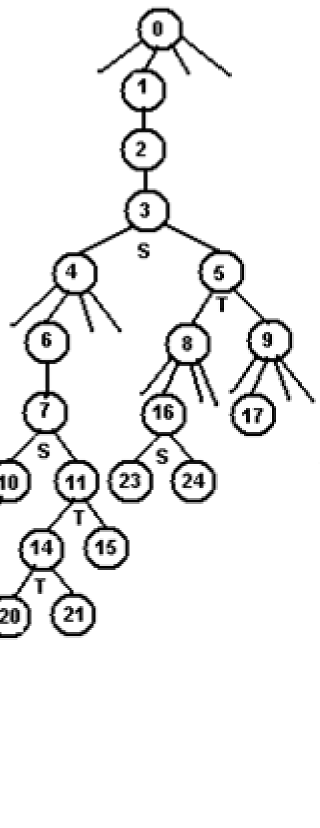
Example: Face Detection



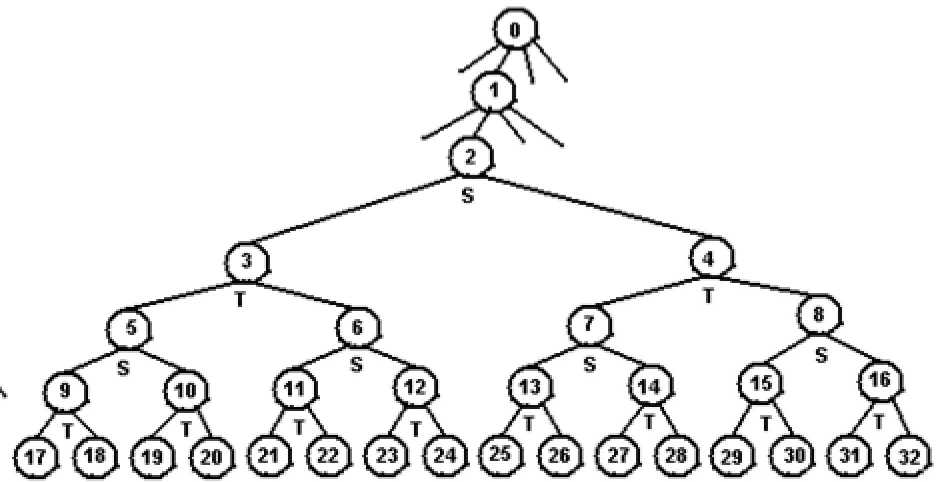
The first two levels of construction. Indicated are false positive rates.

Example: Face Detection (cont)

Learned hierarchy



Manual hierarchy



False positives per image

Detection rate	Manual hierarchy	Learned hierarhcy
92.5%	3.26	1.89
91.1%	1.85	1.02
89.1%	1.11	0.67
85.5%	0.67	0.45
75.5%	0.11	0.07

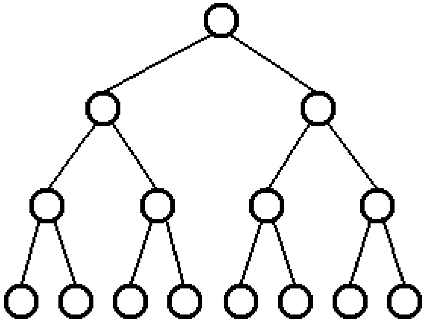
Outline

- Semantic Scene Interpretation
- A Statistical Framework for CTF Classification
  - Part I: Exploring a Hierarchy: “20Q Theory”
  - Part II: Constructing a Hierarchy
  - Part III: Assigning Likelihoods: The “Trace Model”

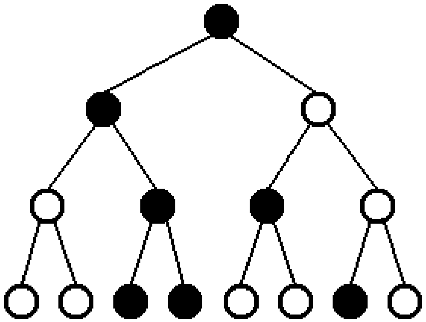
Part III: Trace Model for Assigning Likelihoods to Detections

- Encodes the computational history using a graphical representation
- T: tree underlying the hierarchy
- $S(I)$: subtree of T determined by BFCTF search on image I
- $Z(I) = \{ X_\eta, \eta \in S(I) \}$
- Trace: labeled subtree

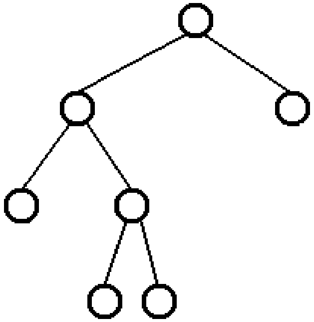
Trace Representation



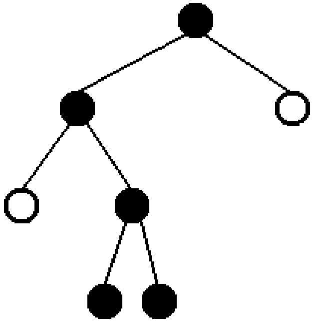
Tree hierarchy



Labeled tree: test responses

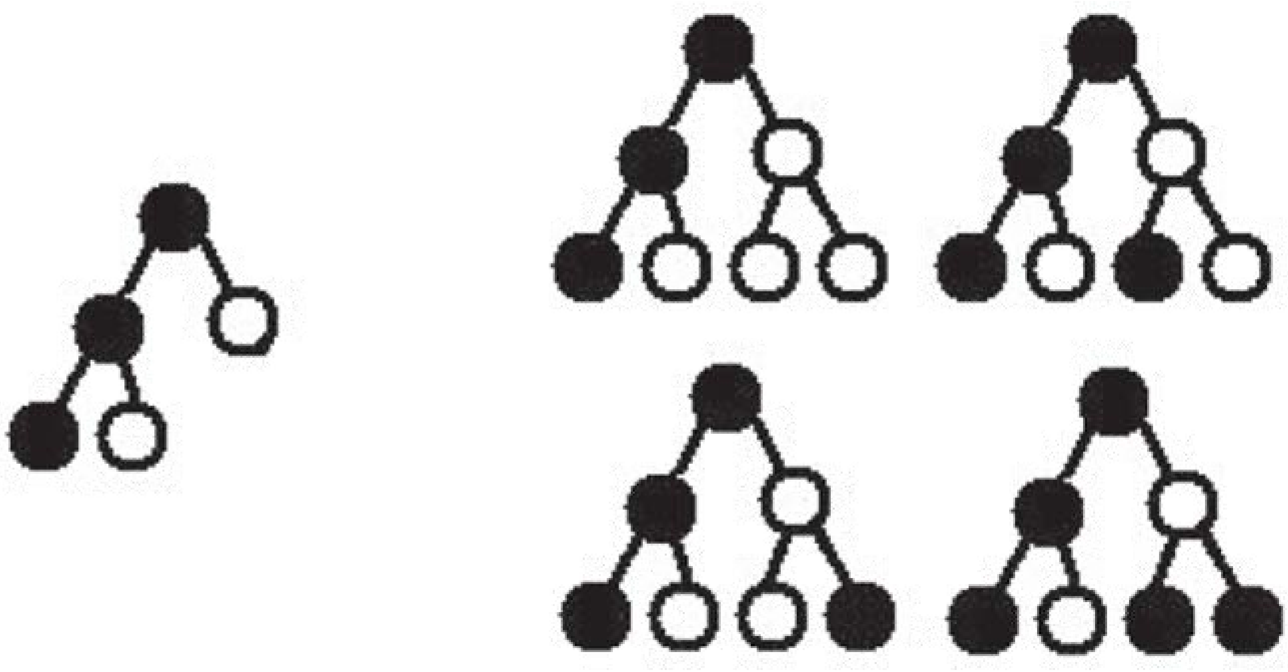


Subtree from BFCTF search



Trace: labeled subtree

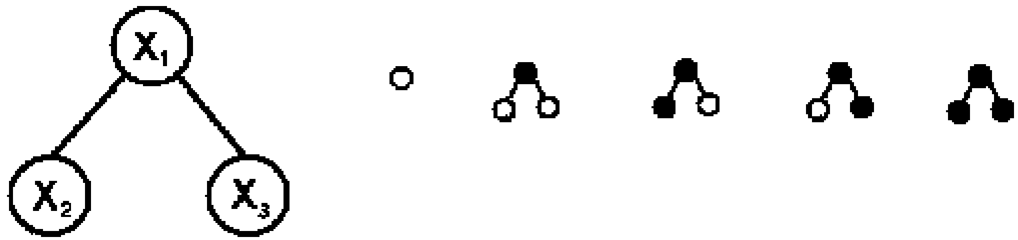
Classifier Realizations to Traces



A single trace produced by four different full tree realizations.

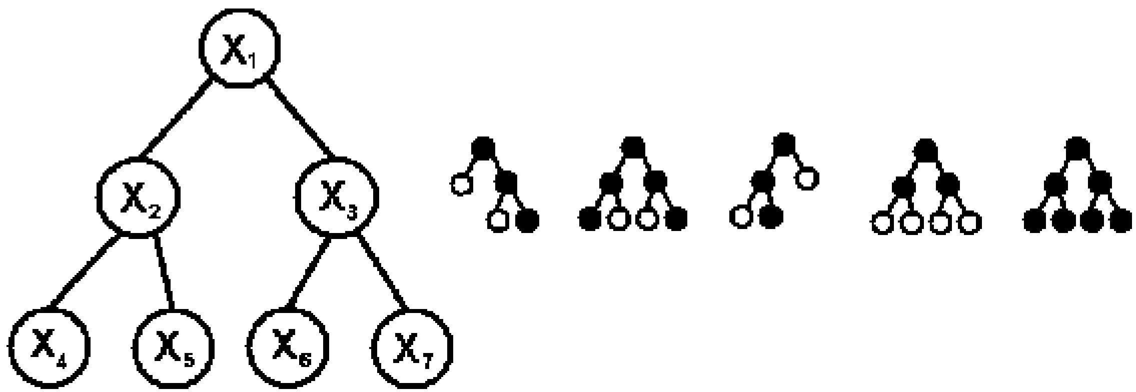
Trace Representation (cont)

$2^3 = 8$



n(T) = 5

$2^7 = 128$



n(T) = 26

Top: A 3 node hierarchy and its 5 possible traces

Bottom: A 7 node hierarchy and 5 of its 26 possible traces

Trace Distributions

The mapping τ : $\boldsymbol{X} \rightarrow Z$, partitions the configuration space:

$$\sum_{z \in \mathcal{Z}} p_{\mathbf{X}}(\tau^{-1}(z)) = 1$$

THEOREM: *Let* $\{p_\eta, \eta \in T\}$ *be any set of numbers with* $0 \leq p_\eta \leq 1$. *Then*

$$P(z) = \prod_{\eta \in S_z} p_\eta(x_\eta)$$

defines a probability distribution on traces where S_z *is the subtree identified with* z *and* $p_\eta(1) = p_\eta$ *and* $p_\eta(0) = 1 - p_\eta$

$$p_\eta(x_\eta) = P(X_\eta = x_\eta | X_\xi = 1, \forall\, \xi \in \mathfrak{A}(\eta))$$

Trace Distributions (cont)

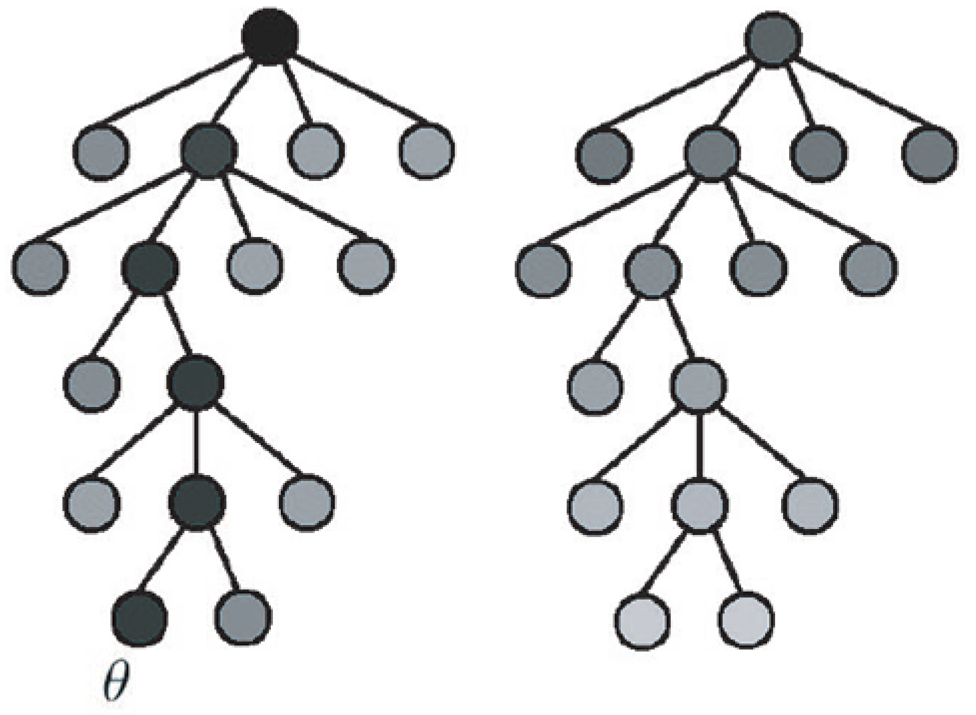
Proof:

- Follows from “peeling” arguments in graphical models
- For a given terminal node, divide the traces into 3 groups:
    - $\eta \notin S$
    - $\eta \in S, x_\eta = 1$
    - $\eta \in S, x_\eta = 0$
- With $p_\eta(1) + p_\eta(0) = 1$ node η is dropped from the summation
- Recursion continues by looping through all the leaves

Application: Face Detection

- *Learning*:
  - *Tests*: Adaboost with binary edge features. Any other learning algorithm could be used as well.
  - *Trace Model*: Learn the probabilities under each interpretation.
- *Interpretations*:
  - *bkg*: represents “no face” (in the subimage)
  - θ_ξ: represents faces with average pose in A_ξ, $\xi \in \partial T$

Estimated Trace Models



Object and background trace parameters: The segment of the full hierarchy that corresponds to the complete chain.

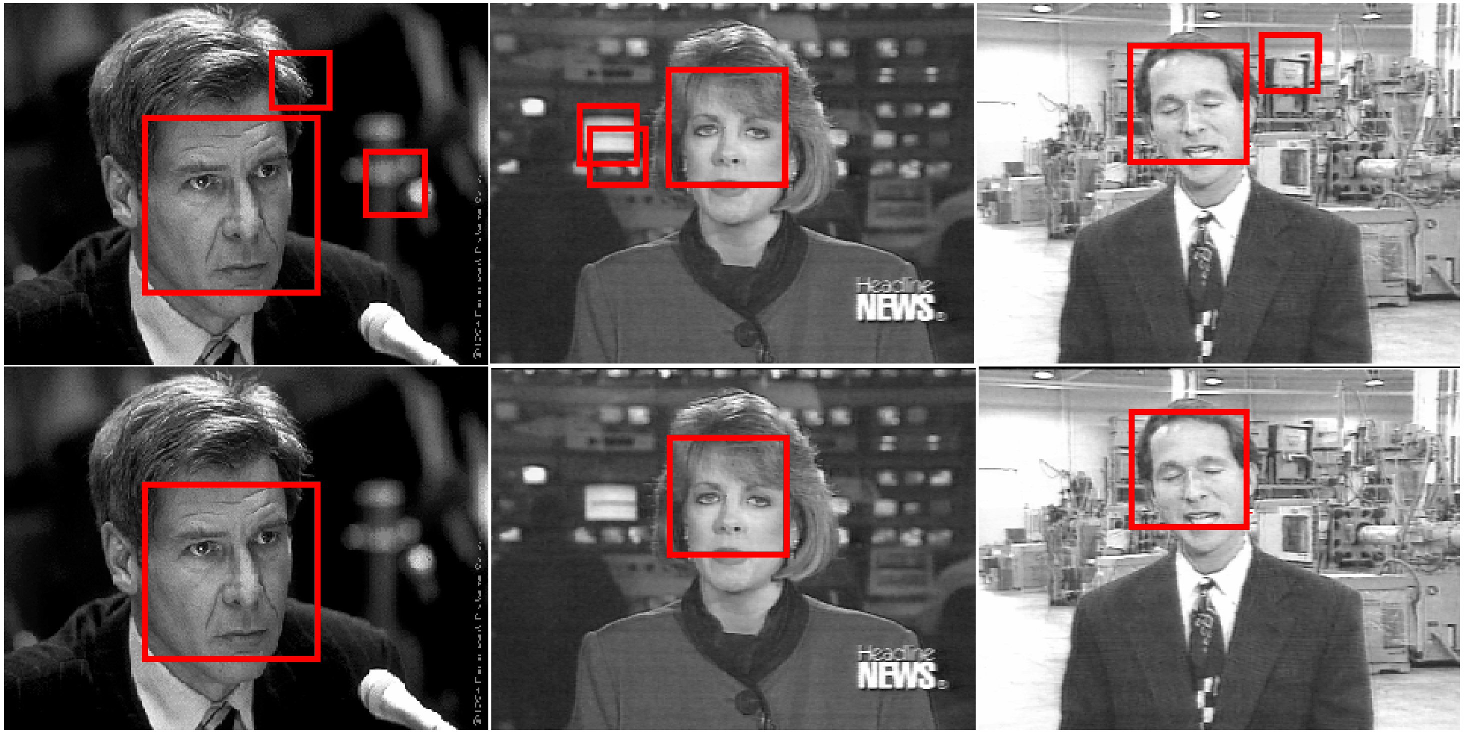
Application: Face Detection (cont)

- Trace-based likelihood ratio test:

$$\frac{P(Z(W)|\theta_\xi)}{P(Z(W)|bkg)} \geq \tau$$

- *Z(W):* trace of image block *W*
- Performed only on complete chains in W
- Requires "learning" of trace models conditional on each pose θ_ξ.

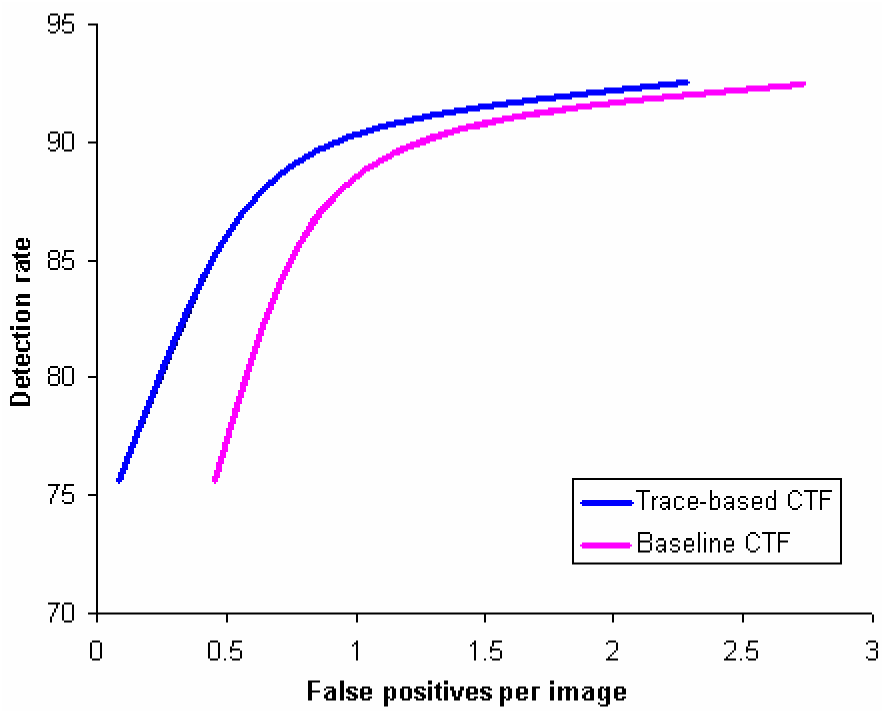
Pruning Detections



Top: Raw results of pure detection

Bottom: False positives are eliminated with the trace model

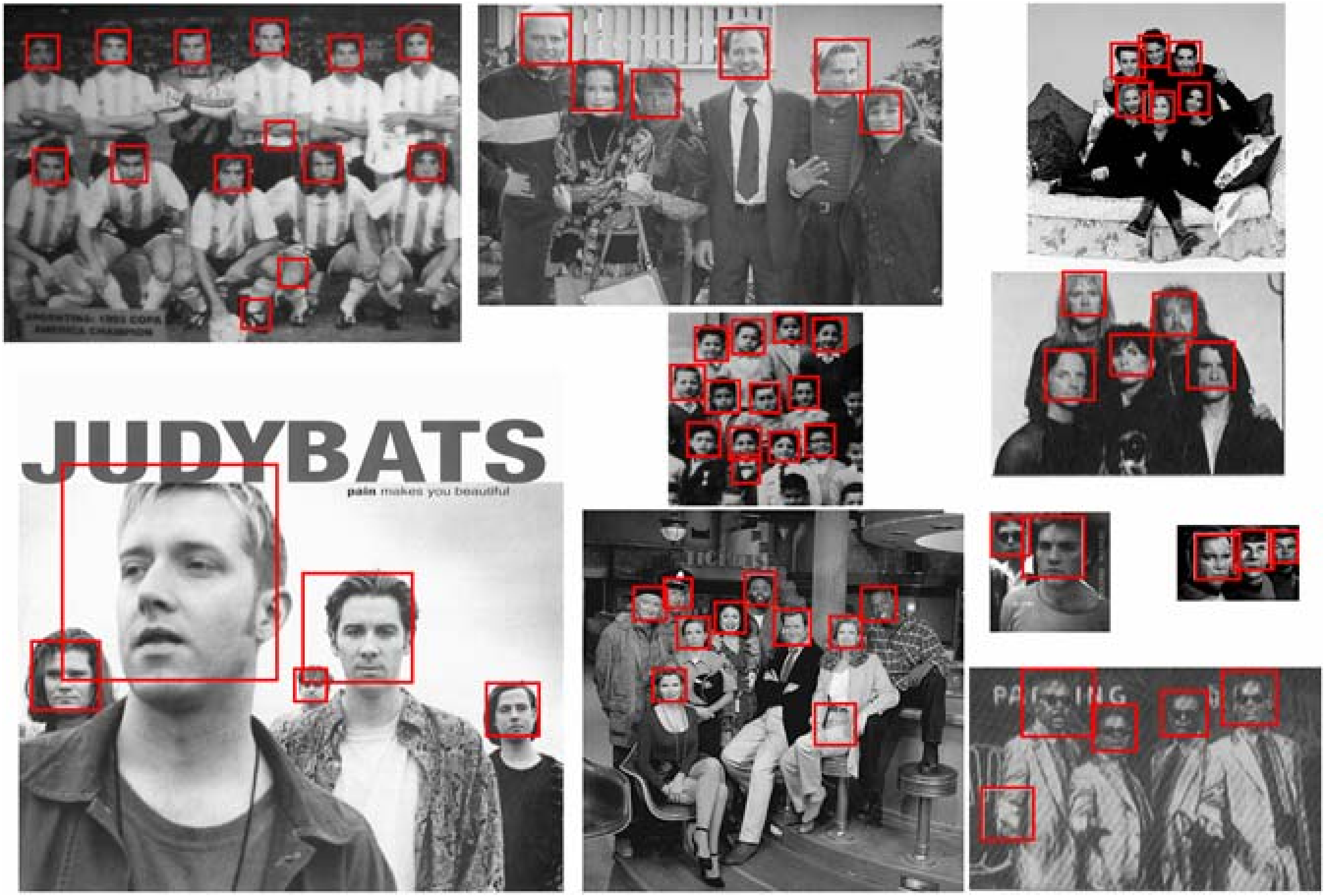
Pruning Detections (cont)



Detection rate vs. false positives on the MIT+CMU test set;

Ex: 0.77 FPs/image at 89.1% detection with |L|=400

Detection Results



Face Tracking

Conclusions

- *Hardwiring efficiency* is a powerful organizing principle.
- Stochastic models on *processing histories* is promising.
- Eventually must test specific hypotheses against specific alternatives.
- Finish the job with rich, contextual models, e.g., *compositional vision*.